Analyzing the Superiority of Logit Standardization in Knowledge Distillation for Efficient Deployment in AAM Environments

1st GwonHan Mun
Mobility Infra Lab
ETRI
Daejeon, South of Korea
gh.mun@etri.re.kr

2nd DaeHo Kim
Mobility Infra Lab
ETRI
Daejeon, South of Korea
daeho@etri.re.kr

***Abstract*—The AAM (Advanced Air Mobility) is an alternative transportation system for passengers and cargo, designed to safely and efficiently serve both urban and rural locations. To safely operate the AAM system, the AAM requires a collision avoidance function using onboard sensors to monitor small aircraft without information transmission capabilities. Among various onboard sensors, camera systems are frequently utilized due to their ability to detect long-range objects using low-cost hardware, it is largely attributed to the advanced algorithms such as deep neural networks (DNNs). However, large DNN models often face challenges in operating efficiently on edge devices, such as mobile phones or other embedded platforms. The knowledge distillation (KD) is a key strategy for reducing network size to meet the demands of industrial applications. However, complex preprocessing schemes for KD hinder an accurate assessment of its standalone performance. To demonstrate the potential of KD in low-cost embedded systems without relying on such schemes, we apply a state-of-the-art KD method with an appropriate batch size strategy to the CIFAR-100 dataset to evaluate its performance.**

***Index Terms*—AAM, CPFSK, GFSK, QPSK, OQPSK, RRC, PAPR, OOB**

I. Introduction

Advanced Air Mobility (AAM) is an emerging alternative transportation system that operates at low altitude, typically between 300 and 600 meters. It requires a higher level of safety than ground-based transportation due the operational characteristics inherent to aeronautical systems. To ensure safe and efficient operation of the AAM system, key components such as vehicle-to-vehicle (V2V) communication system, which transmit an aircraft's status and intentions to surrounding aircraft, and the onboard sensing system for collision avoidance are required.

The onboard sensing system detects and identifies the surrounding object, including small aircraft that lack information transmission capabilities. Among various onboard sensors, camera systems are widely employed because they can detect long-range objects with relatively low-cost hardware, a capability made possible by advanced algorithms such as deep neural networks (DNNs), including model like YOLO, a real-time object detection algorithm known for its speed and accuracy. The advancement of DNN technology has led to significant developments in the fields of computer vision [1], speech recognition [2] and natural language processing [3].

Although powerful network models are used to improve performance, these powerful network models normally require high computational and storage costs, making it challenging for such systems to be applied in industrial domains. Knowledge Distillation (KD), popularized by Hinton [4], addresses this challenge by transferring knowledge from a large, complex model (teacher) to a smaller, more efficient model (student). In this approach, the student model is trained using soft targets generated by the teacher model, enabling high performance while reducing resource requirements. KD is particularly valuable in resource-constrained environments, such as mobile devices or embedded systems.

Numerous studies have been conducted to reduce model size while minimizing the loss in accuracy. These approaches can be broadly categorized into two main types: logit-based methods and intermediate feature-based methods, depending on whether they utilize features extracted from multiple intermediate layers. More specifically, KD can be further refined by incorporating relation-based methods, which aim to capture and transfer the relational information between instances.

Since the introduction of FitNet [5], most subsequent research has focused on feature-based methods due to their superior performance compared to logit-based distillation. However, feature-based approaches generally incur higher computational and storage demands during training. To alleviate this burden, more advanced logit-based algorithms have been developed, aiming to achieve competitive performance while maintaining lower resource requirements.

The Teacher Assistant Knowledge Distillation (TACK) [6] focuses on the logit-based performance worsen when there is a large discrepancy between the teacher and student outputs. Therefore, it reduces this discrepancy by resorting to an additional teaching assistant of moderate model size. And the Knowledge Distillation from a Stronger Teacher (DIST) [7] proposes a relation-based loss to relax the strict requirement of Kullback-Leibler (KL) divergence matching. This approach

defines a distance metric between the softened teacher and student predictions using correlation information, computed either batch-wise or class-wise.

Although these studies utilize logits, their primary focus has been on regularization and indirect comparison. To enhance performance by leveraging the higher-level semantics of logits, several works have revitalized logit-based approaches. For instance, the Decoupled Knowledge Distillation (DKD) [8] method assigns separate weights to target and non-target components, while the Multi-Level Learning Distillation (MLLD) [9] method introduces instance-level, batch-level, and class-level alignment of logit outputs.

In this study, we evaluate the performance of the recently proposed Logit Standardization in Knowledge Distillation method [10]. This approach demonstrates that employing distinct teacher–student temperatures, in combination with sample-wise adaptive temperatures, can yield superior results. Our study highlights its potential for deployment in low-cost embedded systems by systematically tuning commonly used parameters, such as batch size, temperature, and weight coefficients to achieve competitive performance without relying on data augmentation.

II. The logit standardization in knowledge distillation

In classic supervised learning, the student network is trained by penalizing the cross-entropy loss between output of student network and the ground-truth label $\mathbf{y} \in \mathbb{R}^C$, where C is the number of classes.

$$L_{CE}(\mathbf{z}_S, \mathbf{y}) = \sum_{c \in C} \mathbf{y}(c) * log(\mathbf{p}_S(c))$$

where $\mathbf{p}_S = softmax(\mathbf{z}_S)$ represents the predicted probability distribution of the student network, and c is index of classes.

In Knowledge Distillation, popularized by Hinton, Vinyals, and Dean [4], this method tries to train the student network by minimizing not only cross-entropy between the output of student network in softmax layer and ground-truth label, but also the KL divergence loss between two soften predictions obtained from the teacher/student network with a fixed temperature τ in the softmax layer.

$$KL(\mathbf{p}_T || \mathbf{p}_S) = \sum_{c \in C} \mathbf{p}_T(c) log\left(\frac{\mathbf{p}_T(c)}{\mathbf{p}_S(c)}\right)$$

$$\text{where } \mathbf{p}_T = softmax\left(\frac{\mathbf{z}_T}{\tau}\right), \mathbf{p}_S = softmax\left(\frac{\mathbf{z}_S}{\tau}\right)$$

The temperature τ controls the smoothness of the distribution. A lower τ sharpens the distribution, enlarging the difference between two distributions and making the distillation process focus on the maximal logits of the teacher's predictions. In contrast, a higher τ flattens the distribution, narrowing the gap between the distribution and encouraging the distillation process to consider the entire distribution.

Many studies have been conducted to improve performance, and these approaches generally assume that the teacher and student share the same temperature value. However, the logit standardization in knowledge distillation [10] analyzes the effect of shared temperatures and demonstrates that using distinct temperatures for the teacher and student, as well as sample-wise adaptive temperatures, is more effective. This approach aligns with the principle of entropy maximization with a flexible Lagrangian multiplier.

The Logit Standardization in Knowledge Distillation highlights that, in classification tasks, the softmax function is the unique solution for maximizing entropy under the normalization condition of probability and a constraint on the expectation of state in information theory. Extending this derivation, the entropy maximization formulation is applied in the context of KD rather than standard classification. Given a well-trained teacher with prediction $\mathbf{p}_T$, the objective function for student prediction is formulated as follows:

$$\max_{\mathbf{p}_S} L = -\sum_{n=1}^{N}\sum_{c=1}^{C} \mathbf{p}_S^{(n)}(c) log(\mathbf{p}_S^{(n)}(c))$$

$$s.t. \begin{cases} \sum_{c=1}^{C} \mathbf{p}_S^{(n)} = 1, \quad \forall n \\ \sum_{c=1}^{C} \mathbf{z}_S^{(n)}(c)\mathbf{p}_S^{(n)}(c) = \mathbf{z}_S^{(n)}(y_n), \quad \forall n \\ \sum_{c=1}^{C} \mathbf{z}_S^{(n)}(c)\mathbf{p}_S^{(n)}(c) = \sum_{c=1}^{C} \mathbf{z}_S^{(n)}(c)\mathbf{p}_T^{(n)}(c), \quad \forall n \end{cases}$$

Applying the Lagrangian multipliers and taking the partial derivative with respect to $\mathbf{p}_S$, leads to the following solution form by setting the derivative to zero:

$$\mathbf{p}_S^{(n)}(c) = \frac{exp(\beta^{(n)}\mathbf{z}_S^{(n)}(c))}{\sum_{c=1}^{C} exp(\beta^{(n)}\mathbf{z}_S^{(n)}(c))}$$

where $\beta^{(n)}$ is a variable that depends on instance n.

Using the above derivations, this scheme defines general formulation by introducing two parameters μ_S and τ_S:

$$\mathbf{p}_S^{(n)}(c) = \frac{exp((\mathbf{z}_S^{(n)}(c) - \mu_S)/\tau_S)}{\sum_{c=1}^{C} exp((\mathbf{z}_S^{(n)}(c) - \mu_S)/\tau_S)}$$

Assuming a well-distilled student model minimizes the KL-divergence loss, its predicted probability distribution aligns with that of the teacher $i, e., \forall c \in [1, C], \quad \mathbf{p}_T(c) = \mathbf{p}_S(c)$.

Then for arbitrary pair of indices $c1, c2 \in [1, C]$, it can easily lead to

$$\frac{exp((\mathbf{z}_S^{(n)}(c1) - \mu_S)/\tau_S)}{exp((\mathbf{z}_S^{(n)}(c2) - \mu_S)/\tau_S)} = \frac{exp((\mathbf{z}_T^{(n)}(c1) - \mu_T)/\tau_T)}{exp((\mathbf{z}_T^{(n)}(c2) - \mu_T)/\tau_T)}$$

From above equation, the following relation can be derived:

$$(\mathbf{z}_S^{(n)}(c1) - \mathbf{z}_S^{(n)}(c2))/\tau_S = (\mathbf{z}_T^{(n)}(c1) - \mathbf{z}_T^{(n)}(c2))/\tau_T$$

Moreover, by taking a summation across $c2$ from 1 to C, we obtain

$$(\mathbf{z}_S^{(n)}(c) - \bar{\mathbf{z}}_S) = (\mathbf{z}_T^{(n)}(c) - \bar{\mathbf{z}}_T)$$

where $\bar{\mathbf{z}}$ is the mean value of the logits. Subsequently, by summing the squared differences over $c2$ from 1 to C, the following relationship holds:

$$\frac{std(\mathbf{z}_S)^2}{std(\mathbf{z}_T)^2} = \frac{\frac{1}{C}\sum_{c=1}^{C}(\mathbf{z}_S^{(n)}(c) - \bar{\mathbf{z}}_S)^2}{\frac{1}{C}\sum_{c=1}^{C}(\mathbf{z}_T^{(n)}(c) - \bar{\mathbf{z}}_T)^2} = \frac{\tau_S^2}{\tau_T^2}$$

This relationship indicates that for a well-distilled student model, the following conditions should hold:

$$\mathbf{z}_S^{(n)} = \mathbf{z}_T^{(n)} + \Delta^{(n)}, \quad \frac{std(\mathbf{z}_S)}{std(\mathbf{z}_T)} = \frac{\tau_S}{\tau_T}$$

However, due to the gap in model size and capacity, the student may be unable to produce as wide a logit range as the teacher. Consequently, the condition involving $\Delta^{(n)}$ cannot be satisfied, breaking the well-distillation condition. Moreover, when $\tau_S = \tau_T$ (as assumed in previous work), it requires $std(\mathbf{z}_S) = std(\mathbf{z}_T)$ for the well-distillation condition to hold. However, this condition is often violated.

In this paper, to satisfy these conditions, a standardization method for logits is introduced. This method effectively achieves an outcome equivalent to applying instance-specific temperatures, such as $(std(\mathbf{z}_S)\tau), (std(\mathbf{z}_T)\tau)$.

$$\mathbf{p}_S^{(n)}(c) = exp\left(\left((\mathbf{z}_S^{(n)}(c) - \bar{\mathbf{z}}_S^{(n)})/std(\mathbf{z}_S^{(n)})\right)/\tau\right)$$
$$\mathbf{p}_T^{(n)}(c) = exp\left(\left((\mathbf{z}_T^{(n)}(c) - \bar{\mathbf{z}}_T^{(n)})/std(\mathbf{z}_T^{(n)})\right)/\tau\right)$$

This standardization ensures that the relative differences between logits are more significant than their absolute values, thereby enhancing the robustness of the distillation process.

III. Parameter and Simulation Result

Although data augmentation is an effective technique to enhance a model's performance, it hinders the fair evaluation of a rigorous algorithmic performance comparisons. To mitigate such effects and quantitatively evaluate the performance of the recently proposed Logit Standardization in Knowledge Distillation, this study presents a comparative analysis of conventional KD and Logit Standardization in KD for ResNet-8×4 on the CIFAR-100 dataset. For both methods, experiments were conducted with optimally tuned baseline parameters to ensure fair and reliable comparisons. Figures 1 and 2 demonstrate the Logit Standardization in KD achieves a consistent reduction in loss and a corresponding improvement in accuracy compared to the baseline model.

Algorithm 1: Weighted $\mathcal{Z}$-score function

Input: Input matrix $\mathbf{X}$ and τ
Output: Standardized matrix $\mathcal{Z}(\mathbf{X}, \tau)$
Init: $\mathbf{Y} = zeros(B, K)$, *where B: batch size, K: image class number.*
foreach *b = 0 : (B-1)* **do**
  $\mathbf{x} = \mathbf{X}(b, :)$
  $\bar{\mathbf{x}} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}(k)$
  $\sigma(\mathbf{x}) \leftarrow \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mathbf{x}(k) - \bar{\mathbf{x}})}$
  $\mathbf{Y}(b, :) = (\mathbf{x} - \bar{\mathbf{x}})/\sigma(\mathbf{x})/\tau$
Return: $\mathbf{Y}$

Algorithm 2: Weighted $\mathcal{Z}$-score function logit standardization pre-process in knowledge distillation.

Input: Transfer set $\mathcal{D}$ with image-label sample pair $\{(\mathbf{X}_n, \mathbf{y}_n)\}$ with batch size 128, Teacher f_T, Student f_S, Kullback-Leibler Loss $\mathcal{L}_{KL}$, Cross Entropy Loss $\mathcal{L}_{CE}$
Output: Trained student model f_S
Init: Max_accuracy = 0, Max_epoch = 250, Base Temperature $\tau = 4.0$, Loss weight $\alpha = 0.011$ and $\beta = 8.01$
Optimizer: SGD optimizer with a stepwise learning rate schedule, starting at a learning rate of 0.1 and decreasing by a factor of 0.1 at predefined decay points such as 150, 180 and 210.
foreach epoch *in* {1,..., Max_epoch} **do**
  foreach $(\mathbf{X}_n, \mathbf{y}_n)$ *in* $\mathcal{D}$ **do**
    $\mathbf{V}_n \leftarrow f_T(\mathbf{X}_n)$, $\mathbf{Z}_n \leftarrow f_S(\mathbf{X}_n)$
    $q'(\mathbf{Z}_n) \leftarrow$ softmax$(\mathbf{Z}_n, 1.0)$
    $q(\mathbf{V}_n) \leftarrow$ softmax$(\mathcal{Z}(\mathbf{V}_n, \tau))$
    $q(\mathbf{Z}_n) \leftarrow$ softmax$(\mathcal{Z}(\mathbf{Z}_n, \tau))$
    Update f_S towards minimizing $\mathcal{L}_{KD} = \alpha\cdot\mathcal{L}_{CE}(\mathbf{y}_n, q'(\mathbf{Z}_n)) + \beta\cdot\tau^2\cdot\mathcal{L}_{KL}(q(\mathbf{V}_n), q(\mathbf{Z}_n))$
  accuracy = performance of f_S on $\mathcal{D}$
  If: epoch >= Min of predefined decay points
    If: accuracy > Max_accuracy
      Max_accuracy = accuracy
      Save model parameter
  Learning rate update according to schedule using epoch

Loss and accuracy results:

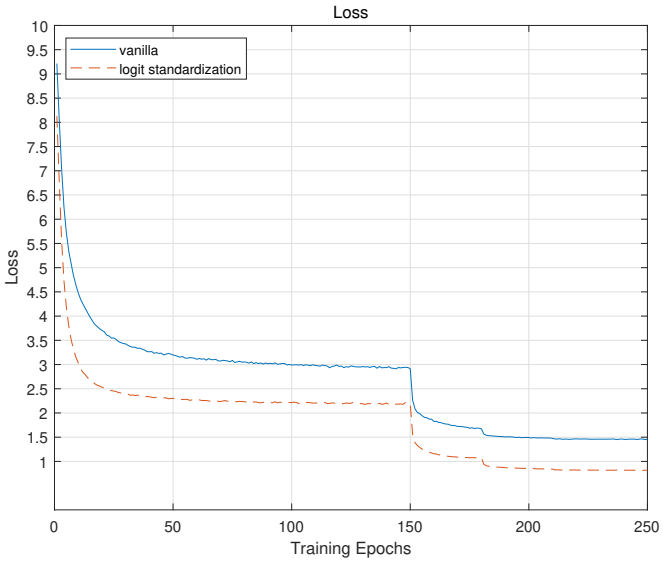

Fig. 1. Comparison of Loss Between Vanilla KD and Logit Standardization KD

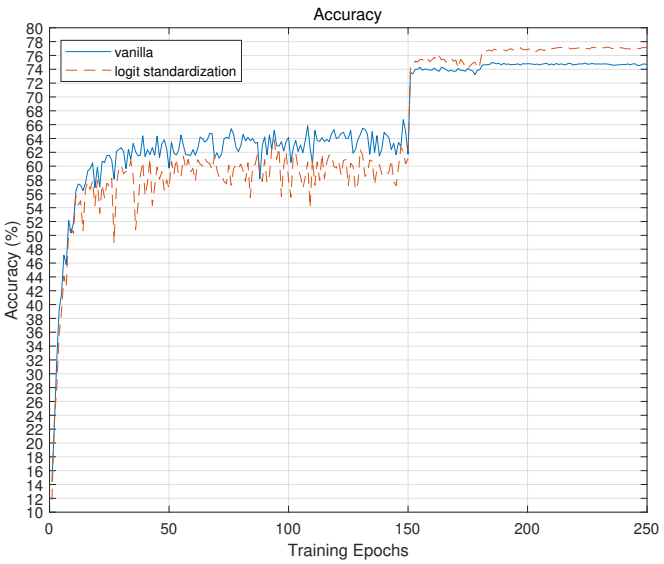

Fig. 2. Comparison of Accuracy Between Vanilla KD and Logit Standardization KD

IV. Conclusion

In this paper, we analyze the effectiveness of the Logit Standardization-based Knowledge Distillation algorithm for efficient onboard implementation. Through optimal parameter tuning and a comparison with conventional KD on the CIFAR-100 dataset without data augmentation, we demonstrate that Logit Standardization consistently delivers more stable and superior performance. This approach supports deployment on AAM platforms with minimal hardware overhead and low cost, thereby improving deployment flexibility and enhancing industrial applicability.

Acknowledgment

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2024-00406112).

References

[1] Huang, G.; Liu, Z.; Maaten, L.; and Weinberger, K. 2017. Densely connected convolutional networks. CVPR 2261–2269.
[2] Chandrashekaran, A.; Kim, J.; and Lane, I. 2017. The capio 2017 conversational speech recognition system. CoRR abs/1801.00059.
[3] Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805.
[4] Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop.
[5] Romero, A.; Ballas, N.; Kahou, S.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.
[6] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In AAAI, 2020.
[7] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. NeurIPS, 2022.
[8] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In CVPR, 2022.
[9] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In CVPR, 2023.
[10] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, Xiaochun Cao. Logit Standardization in Knowledge Distillation. In CVPR, 2024.